Decoding the genomic tree of life

Anne B. Simonson*†, Jacqueline A. Servin‡, Ryan G. Skophammer*, Craig W. Herbold‡, Maria C. Rivera*†, and James A. Lake*†‡§¶

‡Molecular Biology Institute, Departments of *Molecular, Cell, and Developmental Biology and §Human Genetics, and †National Aeronautics and Space Administration Astrobiology Institute, University of California, 242 Boyer Hall, Los Angeles, CA 90095

Genomes hold within them the record of the evolution of life on Earth. But genome fusions and horizontal gene transfer (HGT) seem to have obscured sufficiently the gene sequence record such that it is difficult to reconstruct the phylogenetic tree of life. HGT among prokaryotes is not random, however. Some genes (informational genes) are more difficult to transfer than others (operational genes). Furthermore, environmental, metabolic, and genetic differences among organisms restrict HGT, so that prokaryotes preferentially share genes with other prokaryotes having properties in common, including genome size, genome G+C composition, carbon utilization, oxygen utilization/sensitivity, and temperature optima, further complicating attempts to reconstruct the tree of life. A new method of phylogenetic reconstruction based on gene presence and absence, called conditioned reconstruction, has improved our prospects for reconstructing prokaryotic evolution. It is also able to detect past genome fusions, such as the fusion that appears to have created the first eukaryote. This genome fusion between a deep branching eubacterium, possibly an ancestor of the cyanobacterium and a proteobacterium, with an archaeal eocyte (crenarchaea), appears to be the result of an early symbiosis. Given new tools and new genes from relevant organisms, it should soon be possible to test current and future fusion theories for the origin of eukaryotes and to discover the general outlines of the prokaryotic tree of life.

Today there is enormous interest in discovering the tree of life. But as we get closer to reconstructing it, new experimental and theoretical challenges appear that cause us to reexamine our goals. New obstacles may initially seem insurmountable, but in reality they enrich our understanding of the evolution of life on Earth.

One of the most recent evolutionary mechanisms to challenge our view of genome evolution is the massive horizontal gene transfer (HGT) that has recently become so apparent (1–8). This genetic crosstalk theoretically has the potential to erase much of the history of life that has been recorded in DNA. Indeed, some scientists think that HGT has already effectively erased the phylogenetic history contained within prokaryotic genomes (reviewed in ref. 9).

Although sympathetic to many of these points, we think the best way to decide whether the tree of life is knowable is to try one's hardest to determine it. This article reviews the progress made using whole-genome analyses but does so primarily from the unique perspective of our laboratory.

When Darwin uttered his famous quote, ''The time will come I believe, . . . , when we shall have fairly true genealogical trees of each great kingdom of nature,'' (10) he was not describing prokaryotic life. Rather, he probably envisioned understanding the trees of animal and plant life. In that sense, part of his dream is already a reality. We currently understand the major radiations of the bilateral animals (11, 12), and the relationships linking the major plant groups are starting to be understood (13–17). This review, however, focuses on understanding the radiations that occurred even before those of the plants and animals, namely the enigmatic evolution of prokaryotes and the emergence of eukaryotes.

The origin of the eukaryotes was a milestone in the evolution of life, because eukaryotes are utterly different from prokaryotes in their spatial organization. Eukaryotes, for example, possess an extensive system of internal membranes that traverse the cytoplasm and enclose organelles, including the mitochondrion, chloroplast, and nucleus. This compartmentalization has required a number of unique eukaryotic innovations. The most dramatic innovation is the nucleus, a specific compartment for storing and transcribing DNA, for processing DNA and RNA, and possibly even for translating mRNAs (18). The nucleus is unique to eukaryotes, hence it and the nuclear genome are the defining characters for which eukaryotes are named (eu, good or true; karyote, kernel, as in nucleus).

The prokaryotes, with their simple cellular organization, are generally thought to have preceded the eukaryotes (although see ref. 19). Which prokaryotic groups branched first, however, is not clear, because the root of the tree of life is uncertain and in flux due to a concern that artifacts of phylogenetic reconstruction may have unduly influenced the location of even the root that has the most experimental support (20, 21).

The HGT Revolution

The possibility of analyzing complete genomes awakened interest in prokaryotic genome evolution and profoundly changed our understanding of genome evolution. Before the first genomes were sequenced, there was nearly unanimous scientific agreement that prokaryotic genomes were evolving clonally, or approximately so. In other words, as generation after generation of bacteria divided, each bacterium would contain the DNA it inherited from its parent, except that occasionally a single DNA nucleotide might have mutated, causing a minor change in the daughter genome. Thus it was thought that the family tree derived from any one gene would look like the family tree from any other gene. Diploid eukaryotic cells with two copies of each gene per cell slightly complicated this picture, but they, too, were thought to be evolving clonally. Most researchers felt comfortable with the premise that reliable organismal trees could be calculated from sequences of individual genes. In particular, rRNA genes were favored, because rRNA was easy to sequence, and it was assumed trees calculated from rRNA would probably be the same as those calculated from any other genes. However, it was not acknowledged that HGT had the potential to significantly alter gene trees. For example, if a gene were horizontally transferred from a prokaryote to a human, then the tree

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, ''Systematics and the Origin of Species: On Ernst Mayr's 100th Anniversary,'' held December 16–18, 2004, at the Arnold and Mabel Beckman Center of the National Academies of Science and Engineering in Irvine, CA.

Abbreviations: HGT, horizontal gene transfer; CR, conditioned reconstruction.

¶To whom correspondence should be addressed. E-mail: Lake@mbi.ucla.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0501996102

reconstructed from that gene would place humans in the midst of prokaryotes. Furthermore, each gene tree would show a different set of relationships. (Sometimes one keeps track of whether the transferred genes are new to the genome or whether they replace existing genes. Although this distinction can be important, in this paper, we will refer to both types of exchange as HGT.) Because so much attention was focused on the approximately clonal evolution of rRNA in the pregenomic era, only a few genes other than rRNA were sequenced from multiple organisms, and HGT was largely overlooked.

Once complete genomes were available, the pace of discovery accelerated, as highlighted in early analyses of complete, or nearly complete, genome studies from the laboratories of R. Doolittle (22), W. F. Doolittle (23), Golding (24), Gogarten (25), Ochman (2), and ourselves (8). These and even more recent studies of the evolution of life, based on analyses of complete genomes, described below, revealed the flaws in the old view of clonal evolution. Scientific opinion has now shifted and favors a significant role for HGT in prokaryotic genome evolution.

HGT Has Profoundly Affected Our Understanding of Prokaryotic Genome Evolution

Three remarkable new findings, based on analyses of whole genomes, have engendered appreciation for the important role of HGT in prokaryotic evolution. First, HGT is now generally recognized to be rampant among genomes (rampant at least on a geological timescale). Second, not all genes are equally likely to be horizontally transferred. Informational genes (involved in transcription, translation, and related processes) are rarely transferred, whereas operational genes (involved in amino acid biosynthesis, and numerous other operational activities) are readily transferred. Third, biological and physical factors appear to have altered HGT. These include intracellular structural constraints among proteins (the complexity hypothesis), interactions among organisms, and interactions with the physical environment. These three findings are described below.

Evidence for Extensive HGT

As early as 1996, the complete sequence of the methanogen *Methanococcus janaschii* (26) revealed that its genome consisted of certain groups of genes that were much more similar to eukaryotic genes than those from bacteria, whereas other groups of genes were much more closely related to their bacterial homologs. Koonin *et al.* (27) substantiated that the *M. jannaschii* genes for translation, transcription, replication, and protein secretion were more similar to eukaryotes than to bacteria. They interpreted this finding to mean that archaea were a chimera of eukaryotic and eubacterial genes (27). Using whole-genome phylogenetic methods, our laboratory discovered the presence of two superclasses of genes in prokaryotes that had different relationships to eukaryotic genes. In that study (8) of the *Escherichia coli*, *Synechocystis* PCC6803 (a cyanobacterium), *M. jannaschii*, and *Saccharomyces cerevisiae* genomes (26, 28–30), the *M. jannaschii* informational genes, consisting of gene products responsible for such processes as translation and transcription, were found to be most closely related to those found in eukaryotes. The operational genes of the eukaryote, responsible for the day-to-day operation of the cell (operational genes), on the other hand, were most closely related to their counterparts found in *E. coli* and *Synechocystis* (8). Of the yeast genes analyzed, approximately one-third were informational genes, and two-thirds were operational genes. This provided good evidence that the 16S rRNA tree does not reflect the evolution of all of the genes in a genome and also supplied evidence that early eukaryotes were a chimera of eubacteria and archaebacterial genes. A stylized illustration of these results is shown in Fig. 1. Recently, a thorough comprehensive analysis involving large numbers of genomes and genes has documented the strength of this correlation (31).

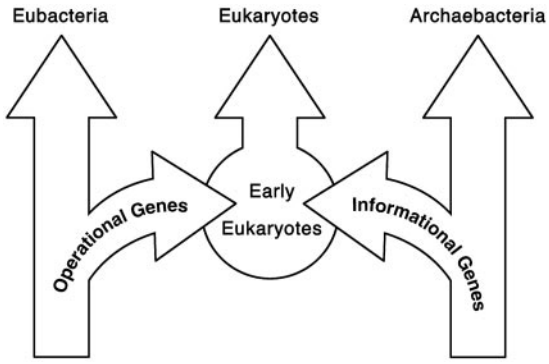

Fig. 1. Early genome studies indicated that eukaryotes were a mixture of eubacterial and archaebacterial genes with an unusual distribution. The operational genes were primarily from the eubacteria, and the informational genes were from the archaebacteria.

Further evidence for extensive HGT came from the observation that another methanogen, *Methanobacterium thermoautotrophicum*, contains several regions that have an ≈10% lower G+C content than the G+C content of the whole genome on average (32). ORFs in these regions exhibit a codon usage pattern atypical of *M. thermoautotrophicum*, suggesting that the DNA sequences may have been acquired by HGT (32).

Additional evidence for HGT came from a thermophilic relative of the methanogens, *Archaeoglobus fulgidus*. ORFs in the functional categories of translation, transcription, replication, and some essential biosynthetic pathways in this prokaryote are very similar to those in *M. jannaschii*. However, these two genomes differ in many of their operational genes, such as those for environmental sensing, transport, and energy metabolism (33). The tryptophan biosynthesis pathway in *A. fulgidus* seems very closely related to the eubacterium *Bacillus subtilis*, even though these two are separated by large distances on the 16S tree (33). These observations suggested that the extent of gene exchange that has occurred in the methanogens and their relatives is tremendous.

Among the extreme thermophiles, some of which live in temperatures in excess of the boiling temperature of water, HGT is equally prevalent (34). Lecompte *et al.* (35) compared the three closely related proteomes from the high-temperature methanogen relatives *Pyrococcus abyssi*, *Pyrococcus furiosus*, and *Pyrococcus horikoshii*. In their gene analysis, the ORFs encoding translation proteins and transcription proteins (informational genes) fairly consistently indicated that the distances among the three species were uniform, as would happen if these genes were evolving approximately clonally. However, most other ORFs (mainly operational genes) gave a wide distribution of distances. The existence of a distribution was interpreted as evidence of HGT (35), because the horizontal transfer of genes from closely and distantly related organisms would be expected to correspond to heterogeneous distances. In addition, *P. furiosus* is capable of transporting and metabolizing maltose/maltodextrin, properties that are absent in *P. horikoshii*. Of two maltose/maltodextrin import systems in *P. furiosus*, one has the greatest similarity to the transport system in *E. coli*, a finding most parsimoniously explained as a lateral transfer of the entire system from *E. coli* to *P. furiosus* (36, 37). Comparison between *P. furiosus* and *P. abyssi* has revealed linkage between restriction-modification genes. Because codon usage is different in various organisms, the codon biases of some restriction-modification systems in the *Pyrococcus* genomes suggest that these systems have been acquired by horizontal transfer (38).

HGT is also widely prevalent in the eubacteria (see the article by H. Ochman, ref. 39); this has been demonstrated in *Aquifex aeolicus*, where little consistency was seen among trees reconstructed from a number of operational genes (40). Comparative analyses of *E. coli* ORFs showed that 675 *E. coli* ORFs have greatest similarity to *Synechocystis*, 231 to *M. jannaschii*, and 254 to the eukaryote *S. cerevisiae* (30). Using skewed base composition and codon usage as a measure of an alien gene, Ochman and coworker (2) argued that 755 of 4,288 *E. coli* ORFs have been horizontally acquired in 234 lateral transfer events, because *E. coli* diverged from *Salmonella* ≈100 million years ago (2).

Classically, the three principal molecular mechanisms known to produce horizontal transfer are transformation, conjugation, and transduction. Numerous authors have found evidence of transduction. For example, the *B. subtilis* genome harbors a number of foreign genes, as evidenced by many prophage-like regions encompassing ≈15% of the genome (41). Like its close relative *B. subtilis*, *Bacillus halodurans,* an alkaliphilic prokaryote, also possesses regions with a G+C content similar to that of some viruses (42). As a consequence of this similarity, those DNA sequences were proposed to have been obtained by lateral transfer (42). The genome of *Clostridium acetobutylicum* contains genes missing in *B. subtilis*. These genes have a number of different phylogenetic relationships. For example, 49 genes reveal an immediate relationship between *C. acetobutylicum* and eukaryotes, and another 195 are most closely related to archaeal extremophiles (43).

The cyanobacterium *Synechocystis PCC6803* is another bacterium whose genome supports extensive HGT among prokaryotes. The genome of *Synechocystis* contains a number of insertion sequence (IS) elements. The DNA in the vicinity of the IS elements displays features of *E. coli* DNA, indicative of horizontal genetic acquisitions (44).

Although HGT Is Rampant, It Is Not Random: The Complexity Hypothesis

In a subsequent phylogenetic analysis (45), our laboratory examined the frequency of horizontal/lateral transfer of operational genes among six prokaryotic proteomes, *E. coli*, *Synechocystis* PCC6803, *B. subtilis*, *A. aeolicus*, *M. jannaschii,* and *A. fulgidus*, using three different topology-based tests of gene ortholog relationships to measure the extent of HGT in informational and operational genes. All three tests showed that operational genes have been continually transferred much more frequently among prokaryotes since the last common ancestor of life or cenancestor (46). To explain at least partially why operational genes undergo HGT more frequently than informational genes, we proposed the complexity hypothesis (45), which posits that informational genes are less likely to undergo horizontal transfer, because their products are members of large complexes with many intricate interactions. Operational genes, on the other hand, are generally not parts of large complexes, and thus are more readily transferred. Obviously the complexity hypothesis is not the sole factor relating differential horizontal transfer rates between informational and operational genes, because many other factors, including environmental ones, can also modify horizontal transfer. At the same time, the data are forcing us to recognize that gene exchange is not simply occurring within species, but extensive exchanges also occur within larger groups of prokaryotes consisting of multiple species as well.

HGT Accelerates Genome Innovation and Evolution

It is becoming clear that HGT has had great impact on the evolution of life on Earth. It is a key agent, perhaps the major agent, responsible for spreading genetic diversity among prokaryotes by moving genes across species boundaries (47). By rapidly introducing newly evolved genes into existing genomes, HGT circumvents the slow step of *ab initio* gene creation and thereby accelerates genome innovation (the acquisition of novel genes by organisms), although not necessarily gene evolution. We refer to a collection of organisms that can share genes by HGT but need not be in physical proximity as an exchange community. In effect, when organisms are exchanging genes, genome innovation is increased in proportion to the effective population sizes of their exchange groups.

We were interested in the structure of exchange communities and in the environmental and other factors that help define them. In an analysis of ≈20,000 genes contained in eight free-living prokaryotic genomes, we assessed which geographic, environmental, and internal parameters have influenced genetic exchange by HGT and found that HGT is not random but depends critically upon these internal and environmental factors. The statistically significant parameters were similar genome sizes, genome G+C compositions, carbon utilization methods, oxygen tolerance, and maximum, optimal, and minimum temperatures (47). By identifying and quantifying those parameters, we were able to delineate exchange community boundaries, estimate the effective population size of exchange groups, and thereby estimate the extent to which HGT has accelerated genome innovation. By correlating the extent of HGT among specific organisms with the degree of phylogenetic clustering of those organisms observed on all possible gene trees, one can determine the effect of various environmental or other parameters on HGT. We found that HGT preferentially occurs among organisms that have environmental and genomic factors in common, a phenomenon we termed positive associativity (47). In short, like prokaryotes preferentially exchanged genes by HGT with like prokaryotes. It is difficult to ascertain precisely how much HGT has accelerated prokaryotic genome innovation, but the acceleration is significant. It has been estimated there are 10^9 prokaryotic species on Earth containing 10^{30} prokaryotes (48). The sizes of exchange communities are unknown, but some of the parameters characterizing them are not too different from those of some terrestrial ecosystems. The median prokaryotic population of 12 diverse soil ecosystem types, as reviewed by Whitman, Coleman, and Wiebe (48), is $\approx 10^{28}$ prokaryotes, suggesting an average exchange group could contain 10^7 species. Allowing 3 orders of magnitude for the inexactness of our estimate, the increase in innovation afforded by HGT could be as small as 10^4, but even this would constitute a huge HGT-dependent increase in innovation. This means that a species exchanging genes only with other members of its species would take 10,000 years to obtain the amount of genome innovation that would occur for an average exchange group in just 1 year. Indeed, HGT may be responsible for a remarkable increase in genome innovation that greatly exceeds anything that could have been accomplished by clonal evolution.

HGT Greatly Complicates Reconstructing the Universal Tree of Life

W. Ford Doolittle recently reviewed the state of "Phylogenetic Classification and the Universal Tree" in a thoughtful analysis (9). He points out the specific challenges to classification that HGT presents as follows, "If, however, different genes give different trees, and there is no fair way to suppress this disagreement, then a species (or phylum) can 'belong' to many genera (or kingdoms) at the same time: There really can be no universal phylogenetic tree of organisms based on such a reduction to genes." In other words, Doolittle (9) suggests that the gene mixing resulting from HGT is so extensive that it might preclude one from ever reconstructing the tree of life. Although it would be disingenuous to pretend that the difficulties are not sizable, our laboratory is pursuing an alternative strategy. We agree that HGT is extensive and imposes limits to phylogenetic reconstruction. However, we also think the only way to discover whether HGT could destroy Darwin's dream of understanding the great kingdoms of nature is to assume that it cannot, and then make

every effort to try to determine the tree of life. Some of the barriers to reconstructing the tree of life and the progress being made to surmount them are discussed below.

Pitfalls in Reconstructing the Tree of Life

Consider what has happened to the once-ebullient field of rRNA phylogenies. For years, phylogenies based on rRNAs were the holy grail of microbial phylogenetics. To be sure, rRNA-based phylogenies have been responsible for many successes, including the new animal phylogeny and demonstrations that the mitochondrion and chloroplast are endosymbionts (11, 12, 49–52). However, prokaryotic phylogenies are another story. One has only to read the latest *Bergey's Manual* (53) to realize that the tree of prokaryotic life is fuzzy and unresolved, so much so that rRNA-based trees, although capable of identifying to which phylum a prokaryote belongs, in most cases cannot determine how the phyla are related to each other. Furthermore, our ability to determine phylogenies accurately depends upon how extensive HGT has been. If very little or no HGT has occurred, then current methods of analysis will allow one to reconstruct the clonal tree of life. At the other extreme, if all genes undergo HGT once per year, then coherent gene trees will be unobtainable. Between these extremes lies a continuum of results, so that perhaps the question we should be asking is, how much phylogenetic information can one obtain, and how can it best be analyzed?

How Can One Reconstruct the Tree of Life in the Presence of HGT?

Presences and absences of genes and gene products have been used for more than two decades to support parsimonious conclusions about the tree of life (54–57). In these analyses, the absences and presences of genes were used as character states, much in the way that nucleotides A, C, G, and T are used as character states in sequence analyses. With the availability of complete genomes, useful methods have been developed for whole-genome analyses (58–61). However, when analyzed using parsimony and simple distance-based methods, these analyses can be significantly influenced by HGT (62, 63).

Recently the prospects of recovering the tree of life in the presence of HGT have improved with the development of a new mathematical algorithm, conditioned reconstruction (CR), for whole-genome-based phylogenetic reconstructions (64). Like some other whole-genome methods, CR analyses also use the absences and presences of genes as character states but, through the use of a reference genome, they can obtain additional information that is not available in other types of analyses. For example, by restricting the analyses to only the genes present in a reference genome R, one can also estimate the number of gene pairs that are missing in both genomes A and B. This is critical information that is not available without the reference genome, and it allows one to use a very general class of mathematical (Markov) models to reconstruct the tree of life.

In CR, the dynamic deletions and insertions of genes that occur during genome evolution, including the insertions introduced by HGT, actually help provide the information needed to reconstruct phylogenetic trees. CR appears to have the potential to reconstruct deeper branchings in the tree of life than is possible with sequence analyses, because whole gene characters evolve more slowly than nucleotides, amino acids, and even gene inserts.

At the same time, it is important to recognize that CRs are not a panacea. It is difficult to assign the gene ortholog sets used by CR analyses accurately, because the process is greatly complicated by the need to distinguish orthologs from paralogs and to simultaneously recognize recently duplicated genes (64). Currently available methods to identify gene ortholog sets are still rudimentary, and new methods are just beginning to be developed. Because CR can be no better than the ortholog sets that it is based on, much improvement is needed in this area.

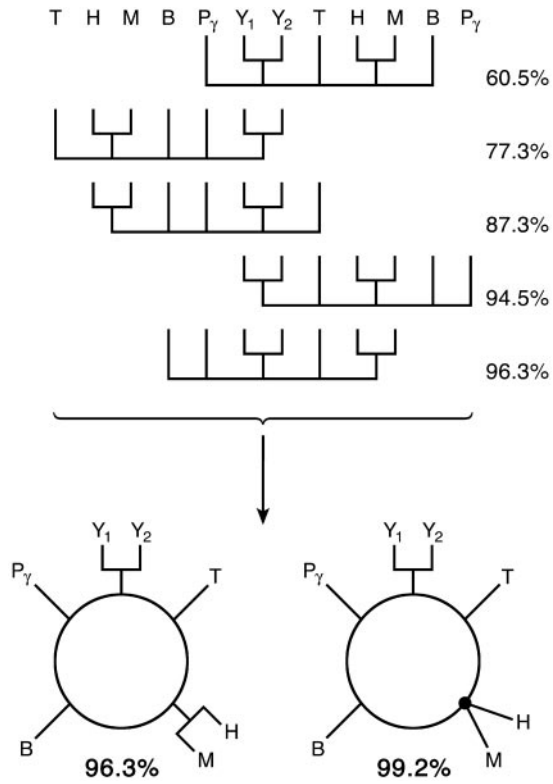

Fig. 2. CRs provide evidence for the ring of life. The genomes are from two yeasts, Y_1 (*Schizosaccharomyces pombe*) and Y_2 (*S. cerevisiae*); a gammaproteobacterium, P_γ (*Xylella fastidiosa*); a bacillus, B (*Staphylococcus aureus* MW2); a halobacterium, H (*Halobacterium* sp. NRC-1); a methanococcus, M (*Methanosarcina mazei* Goe1); an eocyte, T (*Sulfolobus tokodaii*); and an archaeoglobium not shown, the conditioning genome (*A. fulgidus* DSM4304). Cumulative probabilities are shown at the right of each tree. Fully and partially resolved rings are *Lower Left* and *Lower Right*, respectively. [Reproduced with permission from Rivera and Lake (71) (Copyright 2004, Nature Publishing Group).]

Although CR analysis provides a new tool for investigating the tree of life, other methods are also likely to provide important information about deep divergences in the tree of life. These include such important emerging techniques as phylogenetic analyses of concatenated gene sequences (65, 66) or of sets of gene sequences (31, 67), particularly of informational genes, and the analyses of more slowly evolving sequence-related characters such as gene inserts, gene fusions, and even structural domains (68–70). Like CRs, these methods also have their limitations, and much work remains to be done to improve these promising techniques as well.

One of the most remarkable properties of CR is that it can rigorously identify the merger of genomes, a process that until now could not be analyzed using gene sequence. A recently published application of this method has provided evidence that the eukaryotic genome was actually formed by a fusion of the genomes from two disparate prokaryotes.

Evidence That an Ancient Genome Fusion Formed the First Eukaryote

Various theories have been proposed for the origin of the nuclear genes of eukaryotes. These include the autogenous-, chimeric-, and genome-fusion theories. To obtain a better understanding of eukaryotic origins, we analyzed 10 complete genomes using the CR method (71). The sample was comprised of two eukaryotic genomes and eight prokaryotes representing the diversity of prokaryotic life. An additional 24 prokaryotic

genomes were studied in supplementary studies. The results from one analysis are shown in Fig. 2. In this analysis, the five most probable trees are from a set of three Bacteria, three Archaea, and two eukaryotes. The cumulative probabilities of these five trees are shown at the right of each tree. We initially thought that the resolution of the tree was disappointingly poor, because the most probable tree was supported by a low bootstrap value (70% approximately corresponds to the 95% confidence level), and the other trees were supported by even lower values.

However, when the five most probable unrooted trees are aligned by shifting each to the left or the right until their leaves match, they form a repeating pattern indicating that the five trees are simply permutations of an underlying cyclic pattern. (The five most probable unrooted trees are shown with leaves pointing upward to emphasize that each is part of a repeating pattern.) This suggested that they are derived from the single cycle graph (64), or ring, shown in Fig. 2 *Lower Left*. When that ring is cut at any of the five central arcs and then unfolded, the resulting unrooted tree will correspond to one of the five most probable trees. In other words, the data are not tree-like; they are ring-like.

Previously, a combinatorial analysis of the genomic fusion of two organisms had shown that the CR algorithm recovers all permutations of the cycle graph (64). Hence these results can be interpreted in a manner analogous to the interpretation of restriction digests of a circular plasmid or the mapping of a circular chromosome, as implying a ring of life. The fully resolved ring shown in Fig. 2 *Lower Left* is fully consistent with all five of the resolved trees shown in Fig. 2 *Upper*. That ring explains 96.3% of the bootstrap replicates, and the partially resolved ring in Fig. 2 *Lower Right* explains almost all (99.2%) of the bootstrap replicates. These and other control experiments provide robust evidence for the completely resolved ring (Fig. 2 *Lower Left*) and even stronger evidence for the less-resolved ring (Fig. 2 *Lower Right*).

Analyses of this type supported the ring, but other experiments were still necessary to identify the fusion organism. In particular, it was necessary to show that it was the eukaryotes, rather than a prokaryote, that resulted from the genome fusion that closed the ring of life. Hence the identity of the fusion organism was explicitly tested by systematically eliminating the eukaryotes and the individual prokaryotes for the ring of life. The ring opened into a tree only when both eukaryotes were simultaneously deleted from the analysis, indicating the eukaryotic genome had inherited genes from its prokaryotic fusion partners. This then demonstrated that eukaryotes are indeed the products of genome fusions. Furthermore, statistical support for the ring remained high for all possible choices of conditioning genome. From these results and other studies not discussed here, we inferred that the eukaryotic nuclear genome was formed from the genome fusion of either a proteobacterium or a member of a large photosynthetic clade that includes the Cyanobacteria and the Proteobacteria, with an archaeal eocyte as shown schematically in Fig. 3.

Implications of the Ring of Life

Various theories have been proposed for the origin of eukaryotes. These include autogenous, chimeric, and genome fusion theories. The results derived in the CR analyses argue against autogenous theories, i.e., tree of life theories, in which eukaryotes evolved clonally from a single, possibly very ancient, prokaryote. Chimeric theories refer to the acquisition of genes by eukaryotes from multiple sources through unspecified mechanisms. The data presented here argue against them, except of course chimeric theories that specifically propose genome fusions.

At least half a dozen genome fusion theories have been proposed in which the eukaryotic genome originated from two diverse genomes (56, 74–78). These are strongly supported by CR analyses. By default, an endosymbiosis (79) between two prokaryotes is probably the mechanism responsible for the genome fusion observed here, although the fusion signal may have been augmented by gene contributions from eukaryotic organelles. Symbiotic relationships are fairly common among organisms living together and, in rare cases, this leads to endosymbiosis, the intracellular capture of former symbionts (79). Given a genome fusion, and in the absence of other mechanisms that could produce fusions, one concludes that an endosymbiosis was the probable cause.

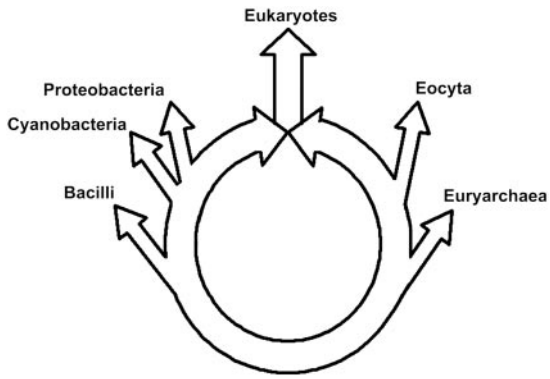

Fig. 3. A schematic diagram of the ring of life. The eukaryotes include all eukaryotes plus the two eukaryotic root organisms, the operational and informational ancestors. Ancestors defining major prokaryotic groups are represented by branching points from the ring. *Archaea* (72), shown on the bottom right, includes the *Euryarchaea*, the *Eocyta*, and the informational eukaryotic ancestor. *Karyota* (73), shown on the upper right of the ring, includes the *Eocyta* and the informational eukaryotic ancestor. The upper left circle includes the *Proteobacteria* (72) and the operational eukaryotic ancestor. The most basal node on the left represents the photosynthetic prokaryotes and the operational eukaryotic ancestor.

Although the data reviewed here solidly support the ring of life, it is important to recognize that CR analysis is a new technique, and its usefulness is still being explored. Currently, the resolution in CR trees is still relatively low. At the same time, it seems unlikely that the ring could be caused by low phylogenetic resolution, because the ring signal monitored in CR analyses is fundamentally different from the parsimony signals that are generated by poorly resolved trees (64).

The ring of life is consistent with and confirms and extends a number of previously reported results. It implies that prokaryotes predate eukaryotes, because two preexisting prokaryotes contributed their genomes to create the first eukaryotic genome. This likely places the root of the ring below the eubacterial– and eocytic–eukaryotic last common ancestors, as shown in Fig. 3. This partial rooting of the ring of life is consistent with the eukaryotic rooting implied by the EF-1α insert that is present in all known eukaryotic and eocytic EF-1α sequences and lacking in all paralogous EF-G sequences (80, 81).

The ring of life also explains some previously confusing observations and raises new ones. Because the eukaryotic genome resulted from a fusion, it is expected that in some gene trees, eukaryotes will be related to *Bacteria*, whereas in other gene trees, eukaryotes will be related to *Archaea*, in accord with the results of others (81–84). The observations of ourselves and others (8, 31), that the informational genes of eukaryotes are primarily derived from *Archaea* and the operational genes are primarily derived from *Bacteria*, are also consistent with the ring. Those observations suggest that the operational genes have come from the eubacterial fusion partner and the informational genes, from the archaeal fusion partner. The ring of life does not explain why the fusion happened, but it provides a broad phylogenetic framework for testing theories for the origin and evolution of the eukaryotic genome. The genome fusion that created the ring of life may in some ways be the ultimate HGT.

We thank M. Kowalczyk for illustrations. This work was supported by grants from the National Science Foundation, the National Aeronautics and Space Administration Astrobiology Institute, the Department of Energy, and the National Institutes of Health (to J.A.L.).

1. Karlin, S., Mrazek, J. & Campbell, A. M. (1997) *J. Bacteriol.* **179,** 3899–3913.
2. Lawrence, J. G. & Ochman, H. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 9413–9417.
3. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002) *Mol. Biol. Evol.* **19,** 2226–2238.
4. Doolittle, W. F. (1999) *Trends Genet.* **15,** M5–M8.
5. Campbell, A. M. (2000) *Theor. Popul. Biol.* **57,** 71–77.
6. Ochman, H. (2001) *Curr. Opin. Genet. Dev.* **11,** 616–619.
7. Koonin, E. V., Makarova, K. S. & Aravind, L. (2001) *Annu. Rev. Microbiol.* **55,** 709–742.
8. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6239–6244.
9. Doolittle, W. F. (1999) *Science* **284,** 2124–2128.
10. Darwin, F. (1887) *The Life and Letters of Charles Darwin* (John Murray, London).
11. Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. & Lake, J. A. (1997) *Nature* **387,** 489–493.
12. Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M. A., Liva, S. M., Hillis, D. M. & Lake, J. A. (1995) *Science* **267,** 1641–1643.
13. Soltis, P. S., Soltis, D. E., Wolf, P. G., Nickrent, D. L., Chaw, S. & Chapman, R. L. (1999) *Mol. Biol. Evol.* **16,** 1774–1784.
14. Nickrent, D. L., Parkinson, C. L., Palmer, J. D. & Duff, R. J. (2000) *Mol. Biol. Evol.* **17,** 1885–1895.
15. Karol, K. G., McCourt, R. M., Cimino, M. T. & Delwiche, C. F. (2001) *Science* **294,** 2351–2353.
16. Pryer, K. M., Schneider, H., Smith, A. R., Cranfill, R., Wolf, P. G., Hunt, J. S. & Sipes, S. D. (2001) *Nature* **409,** 618–622.
17. Pryer, K. M., Schneider, H., Zimmer, E. A. & Banks, J. A. (2002) *Trends Plant Sci.* **7,** 550–554.
18. Hentze, M. W. (2001) *Science* **293,** 1058–1059.
19. Poole, A., Jeffares, D. & Penny, D. (1999) *BioEssays* **21,** 880–889.
20. Philippe, H. & Forterre, P. (1999) *J. Mol. Evol.* **49,** 509–523.
21. Penny, D. & Poole, A. (1999) *Curr. Opin. Genet. Dev.* **9,** 672–677.
22. Doolittle, R. F. & Handy, J. (1998) *Curr. Opin. Genet. Dev.* **8,** 630–636.
23. Brown, J. R. & Doolittle, W. F. (1999) *J. Mol. Evol.* **49,** 485–495.
24. Ribeiro, S. & Golding, G. B. (1998) *Mol. Biol. Evol.* **15,** 779–788.
25. Gogarten, J. P., Murphey, R. D. & Olendzenski, L. (1999) *Biol. Bull.* **196,** 359–361.
26. Bult, C. J., White, O., Olsen, G. J., Zhou, L. X., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al.* (1996) *Science* **273,** 1058–1073.
27. Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997) *Mol. Microbiol.* **25,** 619–637.
28. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., *et al.* (1996) *DNA Res.* **3,** 109–136.
29. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.* (1996) *Science* **274,** 546, 563–567.
30. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., ColladoVides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.* (1997) *Science* **277,** 1453–1462.
31. Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., *et al.* (2004) *Mol. Biol. Evol.* **21,** 1643–1660.
32. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H. M., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., *et al.* (1997) *J. Bacteriol.* **179,** 7135–7155.
33. Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., *et al.* (1997) *Nature* **390,** 364–370.
34. Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I. & Koonin, E. V. (1999) *Genome Res.* **9,** 608–628.
35. Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J. C. & Poch, O. (2001) *Genome Res.* **11,** 981–993.
36. DiRuggiero, J., Dunn, D., Maeder, D. L., Holley-Shanks, R., Chatard, J., Horlacher, R., Robb, F. T., Boos, W. & Weiss, R. B. (2000) *Mol. Microbiol.* **38,** 684–693.
37. Maeder, D. L., Weiss, R. B., Dunn, D. M., Cherry, J. L., Gonzalez, J. M., DiRuggiero, J. & Robb, F. T. (1999) *Genetics* **152,** 1299–1305.
38. Chinen, A., Uchiyama, I. & Kobayashi, I. (2000) *Gene* **259,** 109–121.
39. Ochman, H., Lerat, E. & Daubin, V. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 6595–6599.
40. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., *et al.* (1998) *Nature* **392,** 353–358.
41. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., *et al.* (1997) *Nature* **390,** 249–256.
42. Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N., *et al.* (2000) *Nucleic Acids Res.* **28,** 4317–4331.
43. Nolling, J., Breton, G., Omelchenko, M. V., Makarova, K. S., Zeng, Q. D., Gibson, R., Lee, H. M., Dubois, J., Qiu, D. Y., Hitti, J., *et al.* (2001) *J. Bacteriol.* **183,** 4823–4838.
44. Cassier-Chauvat, C., Poncelet, M. & Chauvat, F. (1997) *Gene* **195,** 257–266.
45. Jain, R., Rivera, M. C. & Lake, J. A. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 3801–3806.
46. Fitch, W. M. & Upper, K. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 759–767.
47. Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. (2003) *Mol. Biol. Evol.* **20,** 1598–1602.
48. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6578–6583.
49. Schwarz, Z. & Kossel, H. (1980) *Nature* **283,** 739–742.
50. Gray, M. W. (1999) *Curr. Opin. Genet. Dev.* **9,** 678–687.
51. Adoutte, A., Balavoine, G., Lartillot, N. & de Rosa, R. (1999) *Trends Genet.* **15,** 104–108.
52. Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B. & de Rosa, R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 4453–4456.
53. Boone, D. R. & Castenholz, R. W. (2001) in *Bergey's Manual of Systematic Bacteriology*, ed. Garrity, G. M. (Springer, New York), Vol. 1.
54. Dickerson, R. E. (1980) in *Diffraction and Related Studies*, ed. Srinivasan, R. (Pergamon, Oxford), Vol. 1, pp. 227–249.
55. Woese, C. R., Pace, N. R. & Olsen, G. J. (1986) *Nature* **320,** 401–402.
56. Lake, J. A., Henderson, E., Clark, M. W. & Matheson, A. T. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 5948–4952.
57. Charlebois, R. L., Singh, R. K., Chan-Weiher, C. C.-Y., Allard, G. C. C., Confaloniere, F., Curtis, B., Duget, M., Erauso, G., Faguy, D., Gaasterland, T., *et al.* (2000) *Genome* **43,** 116–136.
58. Snel, B., Bork, P. & Huynen, M. A. (1999) *Nat. Genet.* **21,** 108–110.
59. Tekaia, F., Lazcano, A. & Dujon, B. (1999) *Genome. Res.* **9,** 550–557.
60. Montague, M. G. & Hutchison, C. A. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 5334–5339.
61. Fitz-Gibbon, S. T. & House, C. H. (1999) *Nucleic Acids Res.* **27,** 4218–4222.
62. Eisen, J. A. (2000) *Curr. Opin. Microbiol.* **3,** 475–480.
63. House, C. H. & Fitz-Gibbon, S. T. (2002) *J. Mol. Evol.* **54,** 539–547.
64. Lake, J. A. & Rivera, M. C. (2004) *Mol. Biol. Evol.* **21,** 681–690.
65. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001) *Nat. Genet.* **28,** 281–285.
66. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. (2000) *Science* **290,** 972–977.
67. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. & Blankenship, R. E. (2002) *Science* **298,** 1616–1620.
68. Gupta, R. S. & Singh, B. (1994) *Curr. Biol.* **4,** 1104–1114.
69. Stechmann, A. & Cavalier-Smith, T. (2002) *Science* **297,** 89–91.
70. Yang, S., Doolittle, R. F. & Bourne, P. E. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 373–378.
71. Rivera, M. C. & Lake, J. A. (2004) *Nature* **431,** 152–155.
72. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
73. Lake, J. A. (1988) *Nature* **331,** 184–186.
74. Gupta, R. S., Aitken, K., Falah, M. & Singh, B. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2895–2899.
75. Martin, W. & Muller, M. (1998) *Nature* **392,** 37–41.
76. Lake, J. A. & Rivera, M. C. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2880–2881.
77. Moreira, D. & Lopez-Garcia, P. (1998) *J. Mol. Evol.* **47,** 517–530.
78. Horiike, T., Hamada, K., Kanaya, S. & Shinozawa, T. (2001) *Nat. Cell Biol.* **3,** 210–214.
79. Margulis, L. (1970) *Origin of the Eukaryotic Cells* (Yale Univ. Press, New Haven, CT).
80. Rivera, M. C. & Lake, J. A. (1992) *Science* **257,** 74–76.
81. Gupta, R. S. (1998) *Microbiol. Mol. Biol. Rev.* **62,** 1435–1491.
82. Brown, J. R. & Doolittle, W. F. (1997) *Microbiol. Mol. Biol. Rev.* **61,** 456–502.
83. Feng, D. F., Cho, G. & Doolittle, R. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 13028–13033.
84. Martin, W., Mustafa, A. Z., Henze, K. & Schnarrenberger, C. (1996) *Plant Mol. Biol.* **32,** 485–491.